2006

Accounting for the uncertainty of information on clustering in the design of a clustered sample

David G. Steel
University of Wollongong, dsteel@uow.edu.au

Robert Graham Clark
University of Wollongong, rclark@uow.edu.au

Accounting for the uncertainty of information on clustering in the design of a clustered sample

Abstract

An important decision that has to be made in developing the design of a cluster or multi-stage sampling scheme is the number of units to select at each stage of selection. For a two-stage design we need to decide the number of units to select from each Primary Sampling Unit (PSU) in the sample. A common approach is to estimate the costs and the variance components associated with each stage of selection and determine an optimal design. This is usually done for estimates of the means or totals of one or a small number of variables. In practice the measure of intra-cluster homogeneity, which is the ratio of the variance components, needs to be estimated from a pilot study or historical data. There may be considerable uncertainty about the intra-cluster correlation. The parameter can be close to zero and the estimate may even not differ significantly from zero, however a design based on zero intra-cluster correlation would be highly clustered and sensitive to any failure of this assumption. This paper considers the effect of uncertainty about the intra-cluster correlation and other relevant population parameters on sample design. We develop an approach to assess this uncertainty using a Bayesian bootstrap method.

Disciplines

Physical Sciences and Mathematics

Publication Details

This conference paper was originally published as Steel, D and Clark, R, Accounting for the uncertainty of information on clustering in the design of a clustered sample, Survey Research Methodology Conference, Taiwan, 2006.

ACCOUNTING FOR THE UNCERTAINTY OF INFORMATION ON CLUSTERING IN THE DESIGN OF A CLUSTERED SAMPLE

David Steel and Robert Clark

Centre for Statistical and Survey Methodology, University of Wollongong, Northfields Avenue, Wollongong NSW 2522, Australia
david_steel@uow.edu.au

ABSTRACT

An important decision that has to be made in developing the design of a cluster or multi-stage sampling scheme is the number of units to select at each stage of selection. For a two-stage design we need to decide the number of units to select from each Primary Sampling Unit (PSU) in the sample. A common approach is to estimate the costs and the variance components associated with each stage of selection and determine an optimal design. This is usually done for estimates of the means or totals of one or a small number of variables. In practice the measure of intra-cluster homogeneity, which is the ratio of the variance components, needs to be estimated from a pilot study or historical data. There may be considerable uncertainty about the intra-cluster correlation. The parameter can be close to zero and the estimate may even not differ significantly from zero, however a design based on zero intra-cluster correlation would be highly clustered and sensitive to any failure of this assumption. This paper considers the effect of uncertainty about the intra-cluster correlation and other relevant population parameters on sample design. We develop an approach to assess this uncertainty using a Bayesian bootstrap method..

Topic Keywords: intra-class correlation, sample design, multi-stage surveys, Bayesian bootstrap

1. INTRODUCTION

An important decision that has to be made in developing the design of a cluster or multi-stage sampling scheme is the number of units to select at each stage of selection. For a two-stage design we need to decide the number of units to select from each Primary Sampling Unit (PSU) in the sample. A common approach is to estimates the costs the variance components associated with each stage of selection and determine an optimal design. This is usually done for estimates of the means or totals on one or a small number of variables. For several two-stage designs commonly in use and assuming a simple

linear cost function the optimal choice is $k = \sqrt{\frac{C_1}{C_2}\frac{1-\delta}{\delta}}$,where C_l is the cost of including a unit at stage l and δ is a measure of homogeneity relevant for the design, which is determined by variance components. In practice the costs and variances for different choices of $\bar{n}$ are calculated as there may be other considerations not reflected in the cost function.

In considering different options it can be useful to examine the design effect (Deff) for the estimates of interest. The Deff of the estimator $\hat{\theta}$ is defined as $V_D(\hat{\theta}) / V_{SRS}(\hat{\theta})$, where $V_D(\hat{\theta})$ is the variance under the design being used and $V_{SRS}(\hat{\theta})$ is the variance under simple random sampling (SRS) (see Kish, 1963, Skinner 1989). Under various assumptions the Deff for a two-stage design can be written as Deff= $\approx 1 + (k-1)\delta$.

In developing the design we need to assume a value for δ. This may come from analysis of census data for a supposedly related variable, data from a previous survey or from a pilot study. In each case there is an element of uncertainty concerning the value of δ. How should this uncertainty be taken into account in choosing the value of k and assessing the likely Deff associated with different choices?

The estimated values of δ are often quite small, perhaps 0.01 or 0.02. A test of the hypothesis that δ=0 may be accepted. In such a case should we proceed on this basis and take as large a cluster size as possible. More generally we may be able to place a confidence interval on δ; we must then decide what value to use. Is the point estimate the best in some sense or should some other approach be used?

In some cases there may be reasons to use quite large values of k. In such cases even small differences in the value of δ assumed can make a large difference to the estimated Deff, which may affect decisions on the total sample size to use.

Similar issues arise with the uncertainty associated with the estimation of the cost ratio.

After the survey has been conducted the inferences will usually take into account the clustered nature of the survey.

2. OPTIMAL DESIGN FOR CLUSTERED SAMPLES

Assume that we are going to select a sample of m Primary Sampling Units (PSUs) and then select a sample of k people or other units from each selected PSU. The total sample size is then $n=mk$ people. For fixed sample size n, sample designs using high values of k are cheaper, but lead to higher standard errors (SEs) for estimates of means and other parameters, whereas using low values of k is more expensive but produces lower SEs.

The sample size should not be fixed but the budget available for the survey will be. Hence we need to consider producing designs that fulfill the cost constriant. Assume a linear cost model:

$$Cost = C = C_0 + C_1 m + C_2 n$$

For many designs used in practice the variance of estimates of mean or total can be expressed as

$$V = V_0 + \frac{V_1}{m} + \frac{V_2}{n} = \frac{A}{n}\left(1 + (k-1)\delta\right)$$

Where $A = V_1 + V_2$ and $\delta = \frac{V_1}{V_1 + V_2}$. This form of variance function applies to expansion, ratio and pos-stratified estimators for sample designs where the first stage units are selected by using either simple random sampling (SRS) or probability proportional to size sampling and the second-stage units using SRS (Hansen et al., 1953). For more complex estimators and designs this form of variance function will be a reasonable approximation. For such designs the likely sampling variance can be calculated using more complex formula.

We can minimise V with respect to *m* and *k* subject to fixed C by using:

$$k = \sqrt{\frac{C_1}{C_2}\frac{1-\delta}{\delta}}.$$

The value of *m* is then determined from the cost constraint.

Often the estimated cost and variances for different choices of *m* and *k* based on are evaluated and other factors are taken into account. Different sample sizes can be used within selected PSUs, but using a constant sample size has practical advantages and is an approach often used in practice.

3. SAMPLE DESIGN USING IMPRECISE INFOMATION

3.1 Sources of Design Information

In practice, *A* and δ must be estimated by some estimators $\hat{A}$ and $\hat{\delta}$. These estimates may be obtained from a pilot survey and therefore be subject to sampling error and possibly biases. They may also be obtained from past surveys with same or similar variable, again introducing issues of sampling error as well as the effect of changes over time, and differences across variables. The professional judgment of the sample designer also comes into play.

The assessment of the sampling variance and the power of analyses of means will be based on $\hat{V} = \frac{\hat{A}}{n}(1 + (k-1)\hat{\delta})$ which, in general, will differ from $V = \frac{A}{n}(1 + (k-1)\hat{\delta})$. These differences may lead to the choice of an inefficient design or and variances larger than planned.

There will also be imprecision associated with the estimation of the cost coefficients and ratio, which we will not consider in this paper.

3.2 Posterior Distributions of Design Effects and Sampling Errors

We want to know the values of the design effect, SEs and power that could occur in the main survey, conditional on what we know from the pilot, which suggests a Bayesian approach. Turner, Prevost & Thompson (2004) derived and estimated posterior distribution of $\delta|\hat{\delta}$ for cluster-randomised trials. However, in general, sample designs involve several complex features. We use the Bayesian bootstrap to deal with all the more complex features in survey design leading to more complex variance formulas than given above.

Rubin(1981) assumed that observations x_1 , …, x_n are independent and identically distributed (i.e.). realizations of a random variable X. Let F be the distribution function of X and let $\hat{F}$ be the empirical distribution function based on these realizations. Let $\theta = f(F)$ be a parameter of interest and let $\hat{\theta} = f\left(\hat{F}\right)$ be an estimator of θ.

Rubin assumed that X was univariate, but this is not necessary for any of the results. We will make use of this in extending the Bayesian bootstrap to two-stage sampling.

Rubin(1981) derived an approximation to the posterior distribution of θ which is similar to the usual frequentist bootstrap of $\hat{\theta}$. The distribution F was assumed to be discrete with support given by the values of X observed in the sample. For simplicity assume that there are no repeated values in the sample although Rubin(1981) did not do this and it is not necessary. Let $p_i = P[X=x_i]$. The unknown parameters consist of $\mathbf{p}=(p_1 , \ldots, p_n)$ which summarise the distribution F, so that $\theta = fn(\mathbf{p})$. The prior distribution of **p** was assumed to be a Dirichlet distribution which is non-informative for **p**. A specific member of the Dirichlet family was assumed where the probability density of **p** is constant for all **p** such that $0 \le \mathbf{p} \le 1$ (i=1,…,n) and $p_1+\ldots+p_n=1$.

Under this prior, the posterior distribution for θ can be approximated by the following procedure:

- Generate (n-1) independent uniform (0,1) random variables $U_1,\ldots,U_{n-1}$.
- Append 0 and 1 to the list these random variables, and sort the variables to give $\{0,U_{(1)},\ldots,U_{(n-1)},1\}$.
- Let $p_1,\ldots,p_n$ be the differences between adjacent values in $\{0,U_{(1)},\ldots,U_{(n-1)},1\}$.
- Generate a sample of size *n* from the distribution with $P[X=x_i] = p_i$ (i=1,…,n).
- Calculate the statistic $\hat{\theta}$ from this sample.

Repeating this process gives the approximate posterior distribution of θ given $\hat{\theta}$. The process is similar to frequentist bootstrapping, because the statistic is calculated from repeated resamples of size *n* from an original sample of size *n*. In frequentist bootstrapping the resamples are selected using simple random sampling with replacement, while in the Bayesian bootstrap the resamples are selected as described above.

The main advantage of the Bayesian bootstrap is that it allows generation of a posterior distribution without fully specifying a parametric model. The main disadvantage is that the choice of prior is somewhat arbitrary. In particular, only values of X which are observed in the sample are used in the posterior distribution, as the support of X is assumed to consist of the observed points only. Rubin (1981) points out that this criticism also applies to the frequentist bootstrap.

When cluster sampling is used where all units in selected PSUs are included, define **X** to be a vector-valued observation which contains all of the data for a cluster. If **X** is defined appropriately, then unequal sized clusters can be accommodated. Then we observe $\mathbf{X_1},\ldots,\mathbf{X_m}$ where *m* is the number of clusters in the pilot sample, and these are assumed to be i.i.d. Clusters can then be resampled using Rubin's Bayesian bootstrap method.

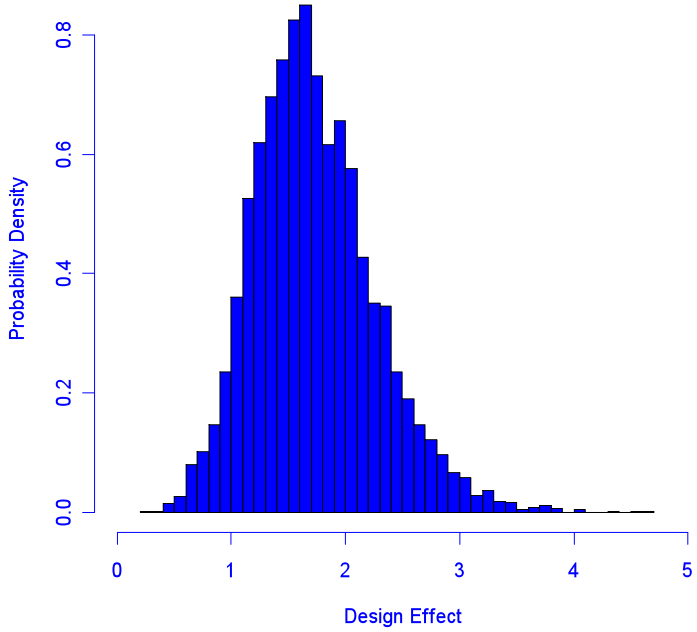
We are investigating the appropriateness of this method when two-stage sampling is used, involving a sample within each PSU.

5. EXAMPLE: DESIGN EFFECTS AND SEs FOR A HEALTH SURVEY

As an example of this approach, consider estimating the design effect and SE for the estimation of number of tobacco smokers in NZ from a health survey. The survey design is assumed to be simple random sampling with replacement of *m* meshblocks, followed by simple random sampling without replacement of *k* people within each meshblock. Meshblocks a re PSUs consisting of an average of 60 people. The design effect and SE for this survey is to be estimated from a pilot survey of 20 meshblocks with sample sizes between 5 and 15 people per meshblock. We used a sample of meshblocks from the previous NZ national health survey to represent the pilot survey.

An unbiased estimator of the design effect (Deff) has been developed using standard techniques This Deff estimator can be regarded as a statistic which is a function of the sample data. The true Deff is the value of this function calculated from the full population data. So in this case θ is the true Deff for the assumed design based on population data and $\hat{\theta}$ is the estimate of the Deff calculated from the pilot survey. The posterior distribution was approximated using the Bayesian bootstrap with 5000 resamples of the meshblocks in the pilot survey. Figure 1 shows the resulting posterior distribution. This figure clearly shows the uncertainty that we have to recongise at the design stage.

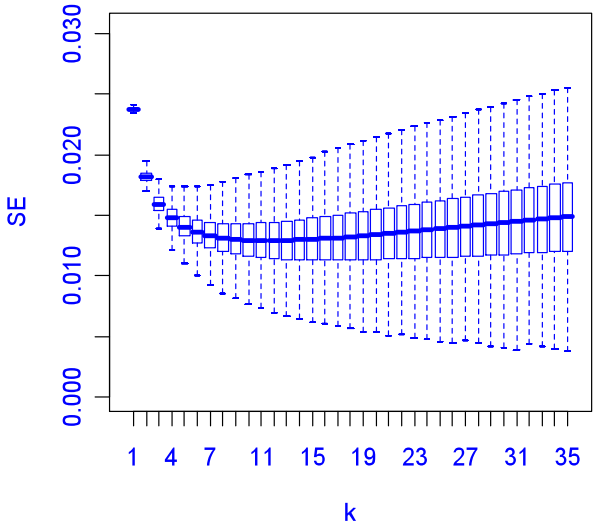
Figure 1: Posterior of Deff for Tobacco (k=10)



The posterior distributions of other parameters can be calculated similarly. For example, Figure 2 shows the posterior distribution of the standard error that will be achieved from the survey for different choices of *k*. A boxplot is shown for this distribution for a number of values of *k*. For each value of *k* and *m* has been recalculated so that the different designs are cost-neutral under the linear cost model where the cost per PSU is 10 times the cost per person in the survey, that is $C_1/C_2 = 10$.

A survey designer can use this could choose which value of *k* is the most appropriate balance of efficiency and robustness. Traditionally a design might be chosen on the basis of the estimated SEs for different values of *k*. Assuming these values are close the posterior mean, then the optimal choice would be *k*=10. Given the shape of the curve in figure 2 for the posterior mean, which has gentler slop for values greater than 10 than for the values less than 10, there might be a case for using a slightly higher value of *k*. However, the ranges shown in figure 2 clearly show that there is greater risk of much higher SEs associated with the higher values of *k*. If we were to choose *k* on the basis upper percentile on of the posterior distribution a choice of *k* around 6 should be considered.

Figure 2: Posterior of SE for different k



3. CONCLUSIONS

In developing sample designs we need to consider the uncertainty in design effect arising from estimation of δ and other parameters that affect the SEs of estimates.

The Bayesian posterior appears a promising approach. The Bayesian bootstrap has several advantages as a tool for understanding the uncertainty in estimates of design effects and other statistics at the sample design stage: It is simple to calculate and does not require a fully specified parametric model which would need to be quite rich to include all of the important features of the population, including variation in PSU size and covariation between PSU size and intra-PSU correlation and the variable of interest. It provides posterior distributions not just confidence intervals. This is useful if we want to look at power averaged over the uncertainty at the design stage.

Our early research suggest that a pilot survey does not provide sufficient information to appreciably help with sample design, unless the pilot sample size is unfeasibly large. As a result, good survey planning really requires some meta-analysis of the design effects and intra-class correlations observed in previous surveys. Uncertainty over time and differences across variables will also be important factors, review of δ and other portable parameters for many situations. This could be reflected in informative priors for intra-class correlation and other parameters. The challenge will be to set up sufficiently realistic models to allow for varying PSU size, and both PPS and SRSR of PSUs, and to build informative priors for these cases. We also need to develop methods to use informative priors in conjunction with the Bayesian boostrap.

REFERENCES

Rubin, D.B. (1981) “The Bayesian Bootstrap,” The Annals of Statistics Vol 9 No. 1, pp.130-134.

Turner, Prevost and Thompson (2004)

Cochran, W. G. (1977). *Sampling Errors and Survey Costs*. New York: Wiley

Hansen, M. H., Hurwitz, W.N. and Madow, W. G. (1953). *Sample Survey Methods and Theory*. New York: Wiley

Kish, L. (1965). *Survey Sampling*. New York: Wiley

Skinner, C. J. (1989).